ELSEVIER

Image and Vision Computing 19 (2001) 807–819

image AND vision COMPUTING

www.elsevier.com/locate/imavis

Face distributions in similarity space under varying head pose ☆

J. Sherrah*, S. Gong, E.J. Ong

Department of Computer Science, Queen Mary and Westfield College, University of London, Mile End Road, London E1 4NS, UK

Received 10 November 1999; revised 25 October 2000; accepted 3 December 2000

Abstract

Real-time identity-independent estimation of head pose from prototype images is a perplexing task requiring pose-invariant face detection. The problem is exacerbated by changes in illumination, identity and facial position. We approach the problem using a view-based statistical learning technique based on similarity of images to prototypes. For this method to be effective, facial images must be transformed in such a way as to emphasise differences in pose while suppressing differences in identity. We investigate appropriate transformations for use with a similarity-to-prototypes philosophy. The results show that orientation-selective Gabor filters enhance differences in pose and that different filter orientations are optimal at different poses. In contrast, principal component analysis (PCA) was found to provide an identity-invariant representation in which similarities can be calculated more robustly. We also investigate the angular resolution at which pose changes can be resolved using our methods. An angular resolution of 10° was found to be sufficiently discriminable at some poses but not at others, while 20° is quite acceptable at most poses. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Gabor filters; Head pose estimation; Similarity representation; Face recognition

1. Introduction

Head pose, closely related to gaze, is an important visual cue for interpretation of human behaviour and intentions. Estimation of head pose from video sequences is a key task for vision systems performing scene understanding for human–computer interfaces or security surveillance. However, the problem is highly complex because the appearance of the face changes with variations in head pose, spatial scale, identity, facial expression and illumination conditions.

While face detection, tracking and recognition have been actively researched for some time, it is usually assumed that the faces are seen at a near-frontal view. The most successful face detection systems that have been developed are based on statistical learning of facial images. Rowley et al. [20] used neural networks to perform face detection for frontal views. Sung and Poggio [21] used a supervised learning approach with a hyper-basis function network to detect faces. Osuna et al. [15] used support vector machines (SVMs) to detect faces. Turk and Pentland [23] used principal component analysis (PCA) to detect faces in their work on EigenFaces. The supervised learning approach based on static views is problematic because correspondence between image points is not explicitly accommodated. Kruger et al. [10] used a deformable graph method to determine face position and pose from learned models. However, this algorithm is iterative and would not be appropriate for real-time applications. Pentland et al. [16] extended the work on EigenFaces to modular EigenSpaces in order to estimate the pose of a face. The concern with this approach is whether sufficient training data can ever be obtained realistically at each pose to establish a reliable PCA basis. Ng and Gong [14] used multiple SVMs for different regions of the pose sphere to perform pose estimation and pose detection across wide pose variations. Li et al. [11] used support vector regression to estimate the pose of a face, then modular SVMs to detect the presence of a face.

2. Motivation and approach

The pose of the head is essentially a three-dimensional quantity being inferred from two-dimensional data, so ambiguities arise. Localisation of the face is implicitly required but intractable when the approximate head pose is not known. Hence the problems of pose-invariant face detection and head pose estimation go hand in hand and must be solved simultaneously.

The task we have undertaken is to develop a system for real-time identity-independent head pose estimation from a

☆ Part of this work was funded by EPSRC ISCANIT project GR/L89624.

* Corresponding author. Tel.: +44-207-975-5230; fax: +44-207-980-6533.

E-mail address: jamie@dcs.qmw.ac.uk (J. Sherrah), sgg@dcs.qmw.ac.uk (S. Gong), ongej@dcs.qmw.ac.uk (E. Ong).

0262-8856/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0262-8856(00)00096-2

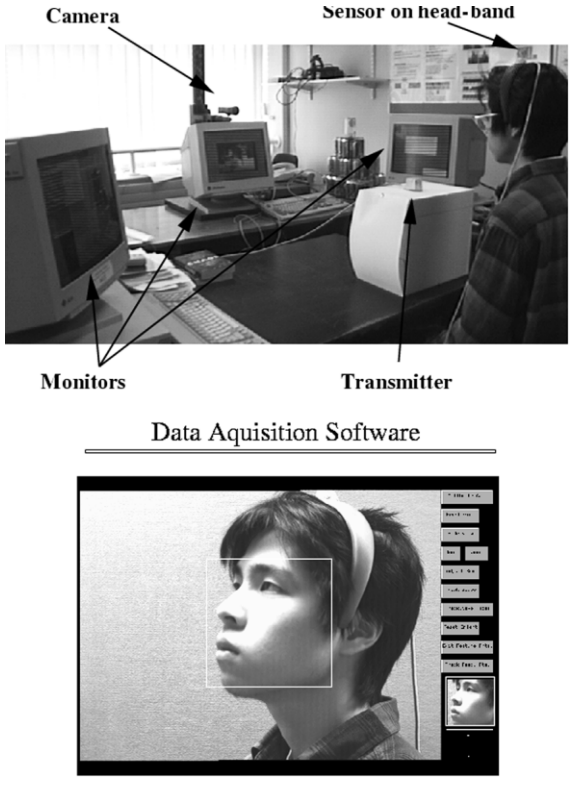

Fig. 1. The system for acquiring labelled views across the view-sphere.

single 2D view. In order to address the enormous problem of simultaneously localising faces and estimating their aspect, two factors are exploited. First, simple attention-focusing cues are used to localise potential facial regions. Efficient focus of attention based upon motion and colour cues has been used to direct face search with a generic appearance-based neural network face detector [12,18]. Second, temporal correlation of head pose and face position are exploited by tracking the face and its pose rather than searching for these parameters directly in each video frame. To associate moving faces in real-time, we adopt a view-based approach that utilises learnable appearance models rather than explicit 3D models. The approach is based on similarities to prototypes [9], which uses second-order similarity to obtain robust similarity measures from sparse data [4]. Therefore neither distributions nor decision boundaries are explicitly calculated. Furthermore, the views are aligned using only simple image-plane transformations such as translation and scaling or at most affine transformation. In particular, no dense correspondences between feature points on different faces are required and as a result, real-time performance is obtained. The approach is similar to work proposed for recognition using similarity measures [2] and for novel view generalisation and synthesis using linear combinations of prototypes [24]. This work extends the idea to wider pose variations and tracking over time.

The method primarily relies on a general assumption that *different people at the same pose look more similar than the same person at different poses*. In other words, pose is a stronger indicator of image–space similarity than identity. The assumption is here referred to as the *pose similarity assumption*. As one can imagine, this assumption is valid only for significant changes in pose. Nevertheless, even for significant pose differences, the assumption may be invalid because intensity images are sensitive to variations in illumination and misalignment. To validate the assumption, the facial images must be transformed to compensate for these variations and to emphasise differences in pose over differences in identity.

The contribution of this work is to experimentally investigate ways of improving performance of the similarity-to-prototypes method for head pose estimation. We investigate the following two issues. First, for a given pose, what transformation of the images is optimal to exaggerate differences in pose and suppress differences in identity? Second, what is the minimum angular separation that can be resolved using similarity-based methods?

The remainder of the paper is laid out as follows. In Section 3, the process of acquiring a training database of labelled face images across the pose sphere is described. A 3D sensor is used to align the data spatially and in pose. In Section 4, the similarity-to-prototypes method is described. Section 5 describes the use of this method for tracking faces under varying pose and identity. It goes on to explain the need for careful choice of image representation in order to make the computation robust. A pose similarity ratio is introduced in Section 6. Two image representations are then examined. First, filtering using Gabor wavelets in Section 7, and second, sub-space compression using PCA in Section 8. The last experimental investigation is in Section 9, examining the smallest angular change in head pose that can be reliably discerned using this method. The conclusions are presented in Section 10.

3. Acquisition of labelled views across the pose-sphere

In order to build appearance models, example views labelled with 3D pose angles (both tilt and yaw) are required. A system was designed that utilises both a magnetic sensor attached to the subject's head and a camera calibrated relative to the transmitter. The sensor was then used to provide pose labels for the face images of the subject captured by the camera. Fig. 1 shows the acquisition system.

More precisely, an electromagnetic 6 DOF Polhemus tracker with a sensor and a transmitter provided 3D co-ordinates and 3D orientation of the sensor relative to the transmitter. The tilt, yaw and roll correspond to rotations about the x, y and z axes, respectively, and are Euler angles.

The sensor was rigidly attached to a head-band worn by the user so that it follows the head's movements and changes in orientation. The image acquisition system used has a single camera, which has been calibrated to the transmitter's co-ordinate system. The location and size of the

Fig. 2. An example labelled head image set. The images of labelled views are from +90° to −90° in yaw and from +30° to −30° in tilt at 10° intervals.

head in the image were determined by back-projection onto the image-plane and an appropriately cropped image is thus acquired. The sensor orientation was used to label the image with head pose.

3.1. Camera calibration

In order to locate and align the 2D head images, camera calibration with respect to the transmitter is needed. This involves determining camera parameters using the 3D positions provided by the sensor and their corresponding 2D projections on the camera's image-plane. Both intrinsic and extrinsic parameters were estimated. The intrinsic parameters are focal length and radial distortion. The extrinsic parameters are the position and orientation of the camera relative to the transmitter's co-ordinate system. We adopted the camera model used by Tsai [22].

The sensor was located in the image by attaching an easily trackable marker to the sensor so that the centre of this marker was at the centre of the sensor. The marker was a solid black circle on white cardboard. The user initialised the tracker by clicking on the marker in the image. The marker was then tracked using a scheme based on intensity thresholding while the user moved it around in 3D space. At regular time intervals, the 3D co-ordinates in sensor space and the 2D co-ordinates in the image-plane were recorded. Typically, a few hundred such data points were recorded. These were then used to perform camera calibration.

3.2. Head alignment and labelling

The position of the sensor with respect to the head is somewhat arbitrary. However, the position and scale of the heads in the images acquired need to be consistent across different people. Therefore, a few facial features were manually located for each subject in order to bootstrap the acquisition process by determining a scaling factor and a 3D point inside the head. This point was rigidly 'attached' to the facial features (eyes and upper lip) and was used to project onto the centre of the acquired head images. In other words, the 3D co-ordinates of facial features were used to determine the co-ordinates of a 3D point inside the head relative to the sensor's 3D co-ordinates. The image was then cropped as determined by the scale factor and re-sampled to a fixed number of pixels.

Acquisition of a subject's facial images proceeds as follows. First the subject's eyes and the middle of the upper lip are located in a frontal view. These features are fairly rigid with respect to the head. Boxes and vertical lines overlaid on the screen help the operator to find the frontal view by assessing bilateral facial symmetry. The inter-ocular line is also required to be horizontal for this view. The distance between the upper lip and the midpoint of the inter-ocular line segment is used to determine the scale factor. The subject is then asked to turn until his face is seen in profile view, i.e. until the 3D orientation estimate indicates rotation through 90°. For the moment assuming that the feature points' depths in the camera co-ordinate system are the same as the depth of the sensor, the three facial feature points project onto epipolar lines in the profile view. They are moved along these lines by the operator until they are at the front of the eyeball. This fixes the z co-ordinate of the eyes with respect to the sensor for the frontal view. The above process can be iterated with feature points being adjusted in frontal and profile view until the operator is satisfied that the 3D facial feature positions have been accurately estimated.

Labelled images were captured with y-axis rotation in the range $\pm 90°$ and x-axis rotation in the range $\pm 30°$ at intervals of 10°. Examples of the captured data for one subject can be seen in Fig. 2.

4. View-based face appearance models using prototypes

Face appearance models are essentially view-based, holistic templates. A simple way to obtain a generic appearance model is to estimate an average face template at each pose. These mean templates can be used to associate face

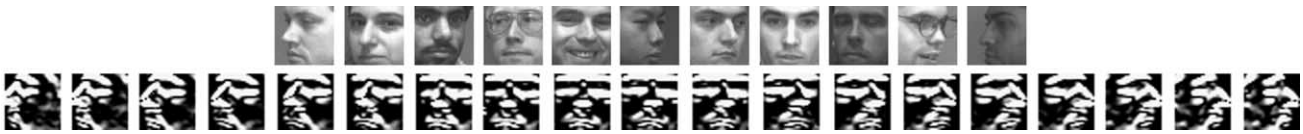

Fig. 3. Top: Example views of 11 different subjects. Bottom: Some of the mean templates obtained by averaging filtered face prototypes at each pose from profile to profile.

images in order to recognise and track poses of faces across viewpoints. Fig. 3 shows some of the mean templates computed by averaging filtered views of 11 different subjects. However, although these view-based mean templates can result in reasonable performance in recognising and tracking pose, they are sensitive to illumination changes and image noise. Furthermore, they do not capture identity information. More elaborate appearance models use linear combinations of training samples. Given sufficient data, such linear combinations can also be statistical models. This includes the use of PCA [16], linear discriminant analysis [5] and hyper-basis function networks [8].

4.1. Linear combination of prototypes

An image $\mathbf{x}$ at a given pose can be decomposed as a linear combination of prototype faces $(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_q)$ at that pose, $\mathbf{x} = \sum_{i=1}^{q} \alpha_i \mathbf{x}_i$. This can be computed using singular value decomposition. The coefficients $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_q)$ minimise:

$$E_{\text{rec}}(\mathbf{x}) = \left\| \mathbf{x} - \sum_{i=1}^{q} \alpha_i \mathbf{x}_i \right\| \tag{1}$$

In the case of linear object classes, the coefficients α are invariant to pose [24]. However, faces do not form a linear class although the approximation is acceptable when pixel-wise correspondence is established. Here we establish no such correspondence due to the need for real-time performance.

4.2. Similarity vectors to prototypes

In order to generalise between views rather than assuming that face appearances are linear combinations of prototypes [24], an image can be represented as a vector of similarities to prototype views [3]. Here we exploit this approach to both face pose tracking and recognition. Let a face image $\mathbf{x}$ at a given pose be represented as a vector $\boldsymbol{\alpha}$ of similarities to q prototype faces $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_q$ at the same pose:

$$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_q], \qquad \alpha_i = h(\mathbf{x}, \mathbf{y}_i) \tag{2}$$

where $i = 1, \ldots, q$ and $h(\cdot)$ is a similarity function that defines a similarity measurement. The calculation is illustrated in Fig. 4.

A straightforward $h(\cdot)$ can be the inverse Euclidean

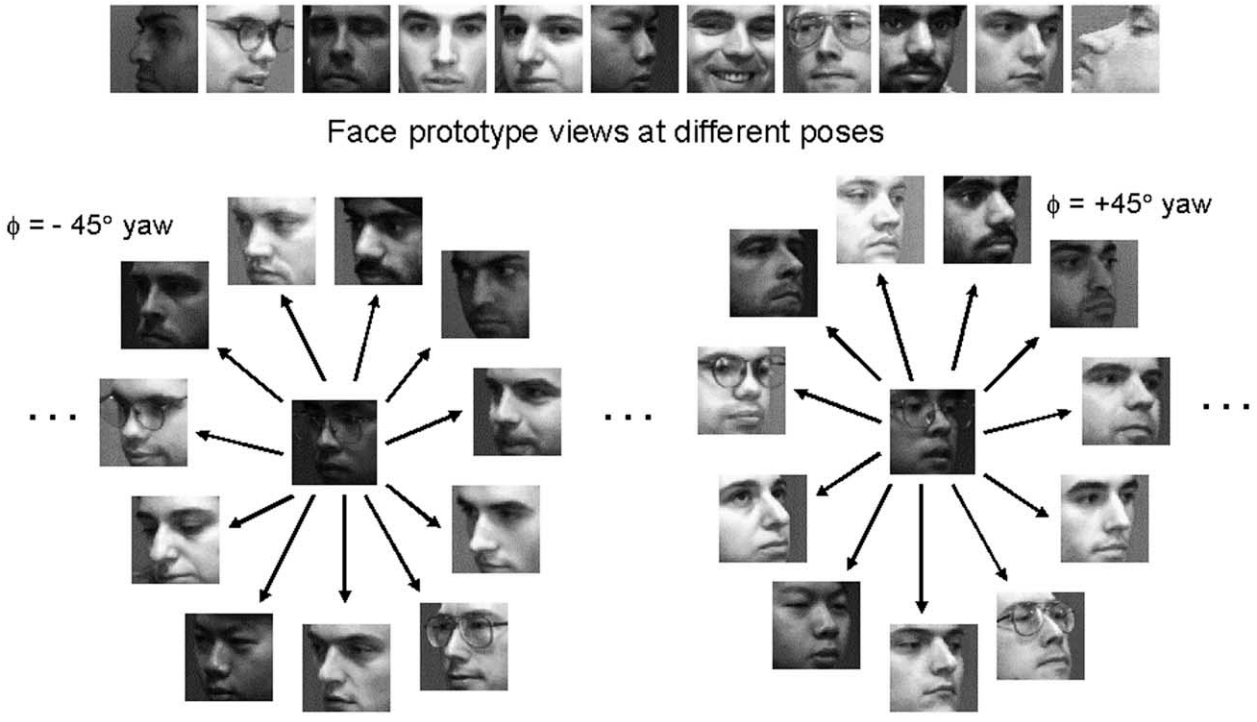

Fig. 4. Illustration of the similarity vector formation process.

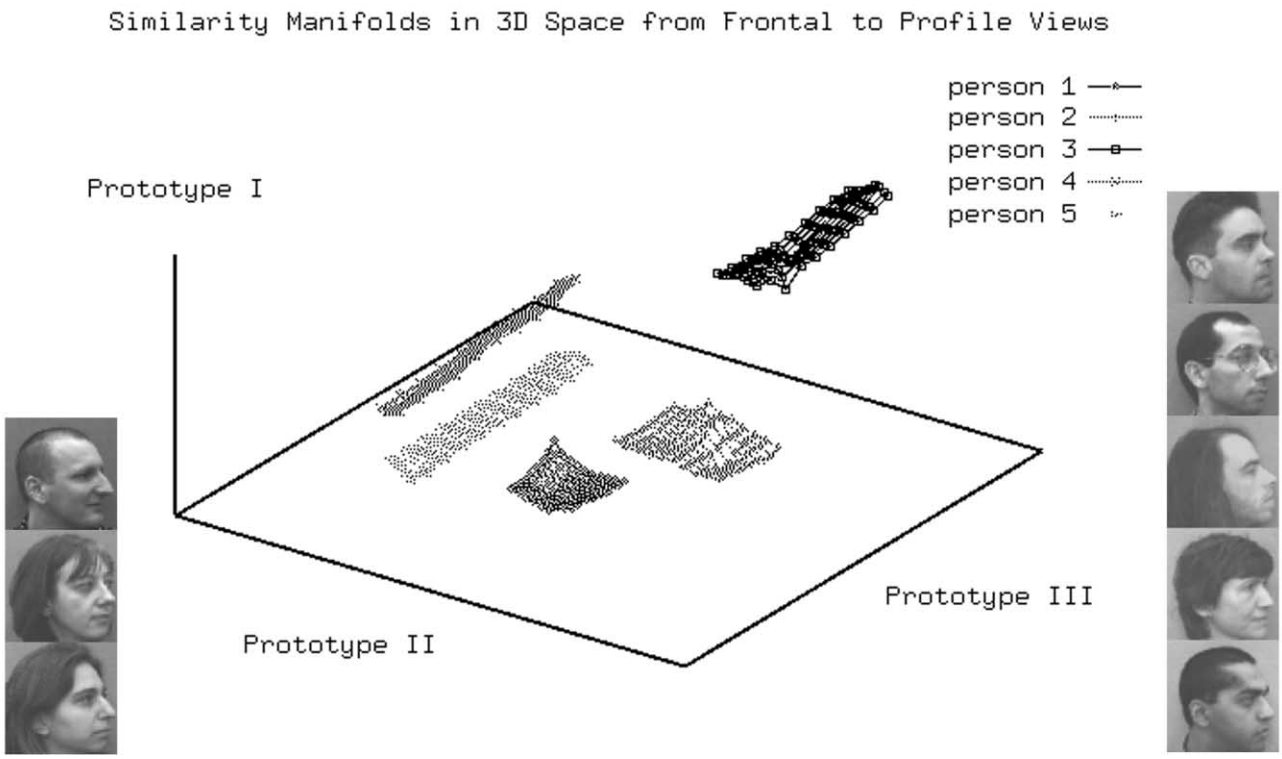

Fig. 5. Left: Prototypes at profile view. Centre: Pose manifolds of novel faces in the vector space of similarity to prototypes. Right: Images of novel faces at profile view.

distance between a face image $\mathbf{x} = [x_1, \ldots, x_N]$ and a prototype $\mathbf{y} = [y_1, \ldots, y_N]$ at a given view where N is the dimensionality of the images:

$$h(\mathbf{x}, \mathbf{y}) = \frac{1}{\|\mathbf{x} - \mathbf{y}\|} = \frac{1}{\sqrt{(x_1 - y_1)^2 + \cdots + (x_N - y_N)^2}} \quad (3)$$

To take normalisation for overall intensity and contrast into consideration, a better measurement should be the inverse of Pearson's linear correlation coefficient [17]:

$$h(\mathbf{x}, \mathbf{y}) = \frac{\sqrt{\sum_{i=1}^{N} (x_i - \mu_{\mathbf{x}})^2} \sqrt{\sum_{i=1}^{N} (y_i - \mu_{\mathbf{y}})^2}}{\sum_{i=1}^{N} (x_i - \mu_{\mathbf{x}})(y_i - \mu_{\mathbf{y}})} \quad (4)$$

where $\mu_{\mathbf{x}}$ and $\mu_{\mathbf{y}}$ are the mean of the elements of $\mathbf{x}$ and $\mathbf{y}$, respectively. Furthermore, a distribution-weighted distance measure such as a Gaussian can also be adopted:

$$h(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2} \right) \quad (5)$$

By measuring similarity vectors of novel faces to prototypes across changes in yaw (y-axis rotation), it can be observed that they often form separable and approximately linear manifolds (see Fig. 5). The model is therefore useful for recognition. Let us, however, consider its use in pose recognition and tracking across views.

5. Transforming facial views to increase within-pose similarity

Given a database of multiple views of different people, a generic view-based appearance model can conceivably be learned for tracking head pose in a person-independent manner given sufficient training data. In practice, the number of examples available at each view is small. Alternatively, appearance models based on similarity vectors to a limited number (in tens) of prototype faces at multiple views can be adopted. Given that face images at the frontal view can be readily detected [12], let a similarity vector $\boldsymbol{\alpha}$ to prototypes for a detected face image at the frontal view be measured using Eq. (2). Pose recognition and tracking can then be performed by finding the next pose θ (both yaw and tilt), which maximises

$$\mathcal{L}(\theta) = \|\boldsymbol{\alpha}_\theta^t\| + \kappa h(\boldsymbol{\alpha}_\theta^t, \boldsymbol{\alpha}^{t-1}) \quad (6)$$

where $\|\boldsymbol{\alpha}_\theta^t\|$ is the L_2 norm of the similarity vector at pose θ at time t. Function $h(\boldsymbol{\alpha}_\theta^t, \boldsymbol{\alpha}^{t-1})$ is the similarity measure between the two similarity vectors at the previously known pose and the currently likely pose. Maximising $\mathcal{L}(\theta)$ imposes two constraints. The first term maximises

the magnitude of similarity regardless of identity in a neighbourhood centred at the likely pose at time t, therefore performing a generic face matching at the likely pose at time t. The second term assumes identity constancy in similarity vector space provided that all other sources of variation such as lighting and translational shift in the images have been eliminated (as shown in Fig. 5). The constant κ controls a trade-off between the two factors and its value will depend on the expected smoothness in the pose change and the variation in a face's similarity measures to prototypes in different views.

Crucially, such a model is based on the assumption that different faces at the same pose are more similar to each other than the same face at different poses. Since pixel-wise correspondence between images is not currently possible for real-time pose estimation, the image data must be transformed to a space in which this assumption is true on average for the chosen similarity criterion.

The most obvious transformation for images is to apply an image filter. The optimal filtering of prototype images is expected to be different at each pose angle because different features are important at different poses [7]. The most natural filtering of images for this task is to use orientation-selective features. Gabor filters are particularly appropriate because they incorporate smoothing, which reduces sensitivity to spatial misalignments. Recent studies on Gabor filters have shown that these filters are approximately the basis functions for natural images [19] and have been discovered in the early visual system of mammals [26].

While filtering may enhance pose-specific features, it is expected to provide only small invariance to identity. Intuitively, a representation of the images is required that encodes only very coarse-scale intensity variations with pose. It has been shown in Ref. [7] that PCA can be used to discard identity information while maintaining pose information. PCA has the extra advantage that similarity measures in a low-dimensional space are more robust and easier to compute than in a high-dimensional space.

In this work, we define a criterion to quantify the goodness of a given transformation method for pose prototypes. The criterion is then used in a series of experiments. In the first experiment, Gabor filters are examined as a method for enhancing pose differences at each pose angle. In the second experiment, PCA is used to represent prototypes and its identity-invariant properties are examined. In the third experiment, the criterion is used to determine the angular resolution at which neighbouring poses can be resolved.

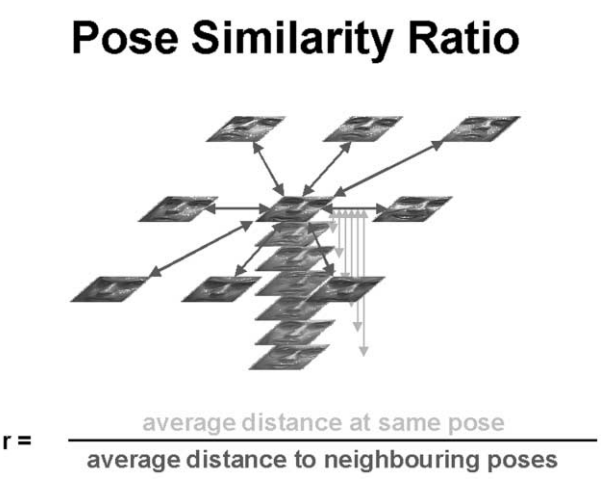

Fig. 6. Illustration of the pose similarity ratio.

6. The pose similarity ratio

When matching images from various poses to a group of prototype images, it is desirable to calculate similarity in a space that is invariant to identity and sensitive to differences in pose. To select a good transformation, a criterion is required to allow us to compare image representations. The criterion should be based on the pose similarity assumption that differences in pose are more significant than differences in identity. Our criterion is defined as the following ratio:

$$r(\phi, \theta, f(\cdot)) = \frac{\bar{d}(\phi, \theta, f(\cdot))}{\bar{d}(\phi \pm \delta\phi, \theta \pm \delta\theta, f(\cdot))} \tag{7}$$

This ratio shall be referred to as the *pose similarity ratio* where:

- $f(\cdot)$ is a transformation function that maps the images to some other representation either with the same dimensionality, e.g. an image filter, or with lower dimensionality, e.g. linear projection;
- $\bar{d}(\phi, \theta, f(\cdot))$ is the average distance (inverse of similarity) between f-transformed prototypes of varying identity at a

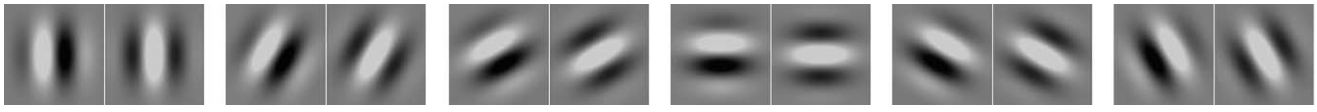

Fig. 7. Gabor filters at different orientations γ. The real part is on the left, imaginary on the right.

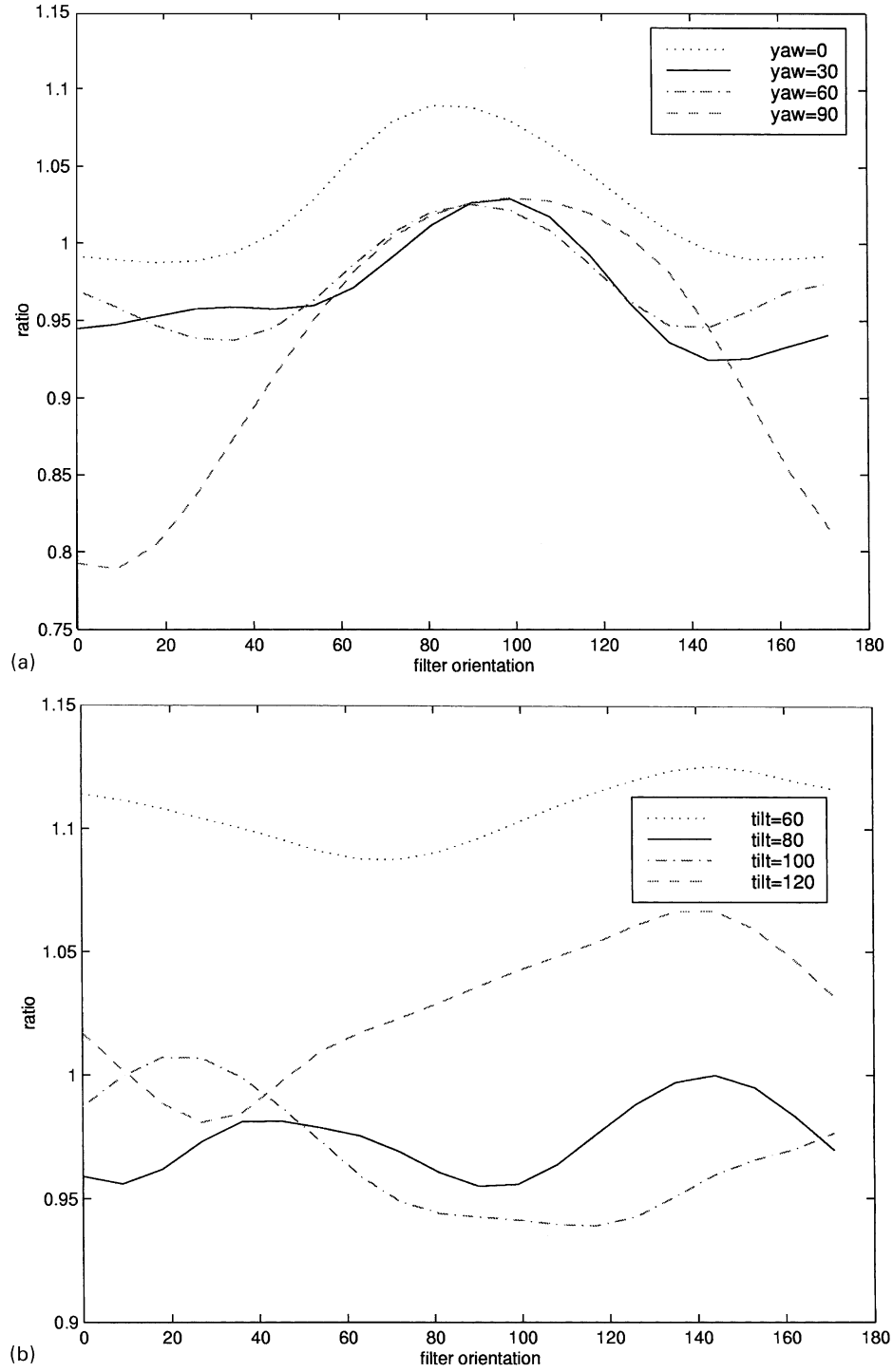

Fig. 8. Pose similarity ratios for varying head pose and filter orientation. (a) Varying yaw with tilt fixed at 90°. The neighbourhood is based on yaw only. (b) Varying tilt with yaw fixed at 90°. The neighbourhood is based on tilt only.

given pose:

$$\bar{d}(\phi, \theta, f(\cdot)) = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d(f(\mathbf{x}^i_{\phi,\theta}), f(\mathbf{x}^j_{\phi,\theta}))}{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} 1} \quad (8)$$

where $\mathbf{x}^i_{\phi,\theta}$ is the prototype image of subject i at pose angles (ϕ,θ) and $d(\mathbf{x}_1,\mathbf{x}_2)$ is the distance between two points in high-dimensional space:

- $\bar{d}(\phi \pm \delta\phi, \theta \pm \delta\theta, f(\cdot))$ is the average distance between f-transformed prototypes at the given pose and prototypes of varying identity and pose over the given range of

Table 1
Average minimum pose similarity ratios for filters of different sizes

Filter size (in pixels)	9	11	13	15
Average best ratio	0.974617	0.964791	0.958090	0.953220

neighbouring poses:

$$\bar{d}(\phi \pm \delta\phi, \theta \pm \delta\theta, f(\cdot)) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{y=\phi-\delta\phi}^{y=\phi+\delta\phi}\sum_{t=\theta-\delta\theta}^{t=\theta+\delta\theta} d(f(\mathbf{x}_{y,t}^{i}), f(\mathbf{x}_{y,t}^{j}))\cdot\delta(y-\phi, t-\theta)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{y=\phi-\delta\phi}^{y=\phi+\delta\phi}\sum_{t=\theta-\delta\theta}^{t=\theta+\delta\theta}\delta(y-\phi, t-\theta)} \tag{9}$$

where $\delta\phi$ and $\delta\theta$ are the sizes of the yaw and tilt neighbourhoods and $\delta(y-\phi, t-\theta)$ is a delta function to discount the distance of a prototype to itself:

$$\delta(a,b) = \begin{cases} 0 & \text{if } a = 0 \text{ and } b = 0; \\ 1 & \text{otherwise} \end{cases} \tag{10}$$

An illustration of the similarity ratio is shown in Fig. 6. The ratio can be interpreted as follows: when the ratio is small, faces at the given pose are more similar to each other than to faces at neighbouring poses and the pose similarity assumption is valid. For large ratio values, faces at neighbouring poses are more similar than at the same pose and the assumption is invalid. At a given pose, the ratio can be minimised with respect to $f(\cdot)$.

We now describe three experiments using the ratio criterion. All results are based on a database of 30×30 images collected from $N = 8$ subjects at poses over the pose sphere of range $\phi \in [0°, 10°, \ldots, 180°]$ and $\theta \in [60°, 70°, \ldots, 120°]$. In all experiments, the distance function used was Euclidean

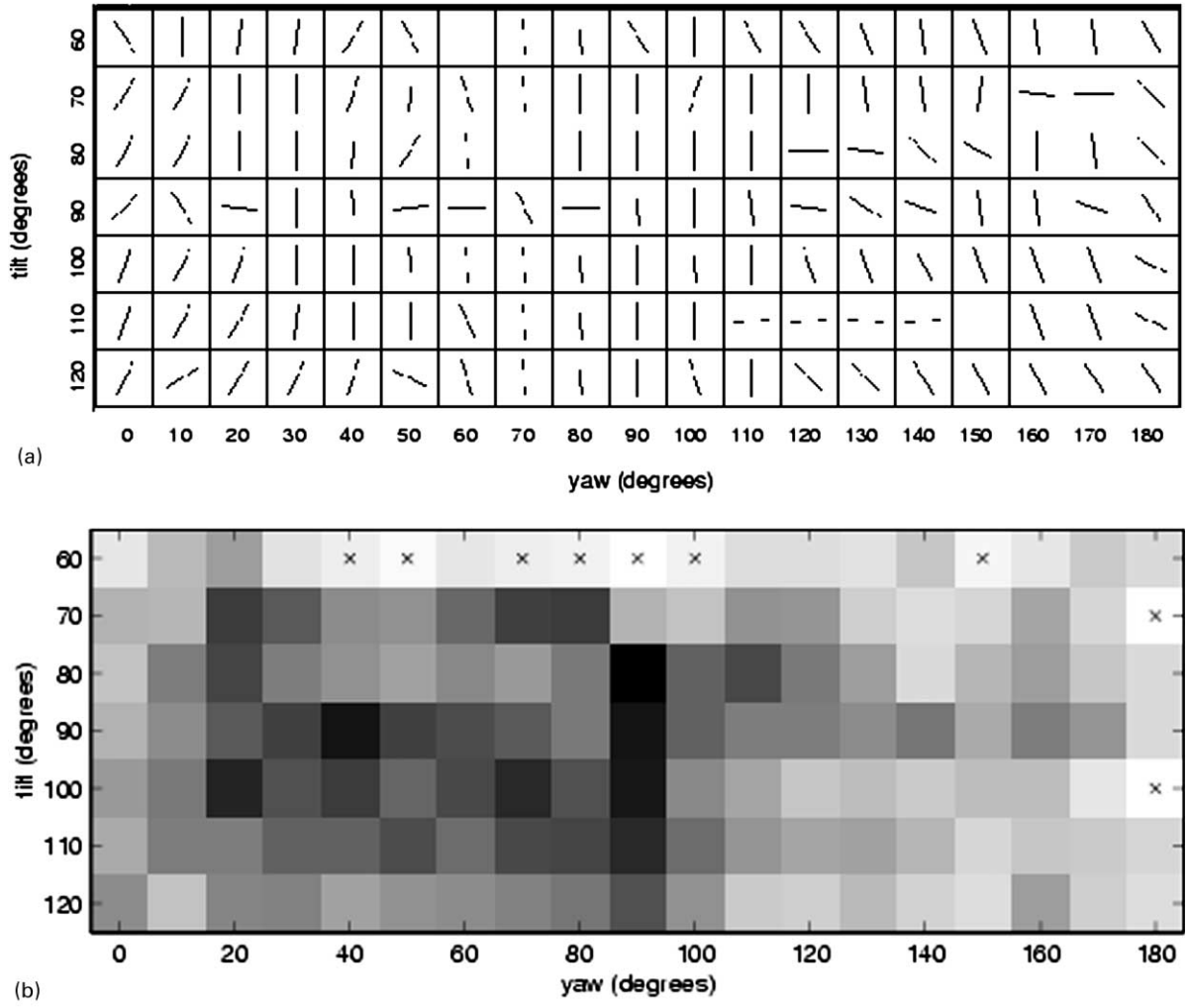

Fig. 9. Results for best filters of size 13×13. (a) Orientations of optimal filters over the pose range. (b) Corresponding minimum pose similarity ratios with whiter cells corresponding to higher ratios. Ratios greater than one are denoted by '×'.

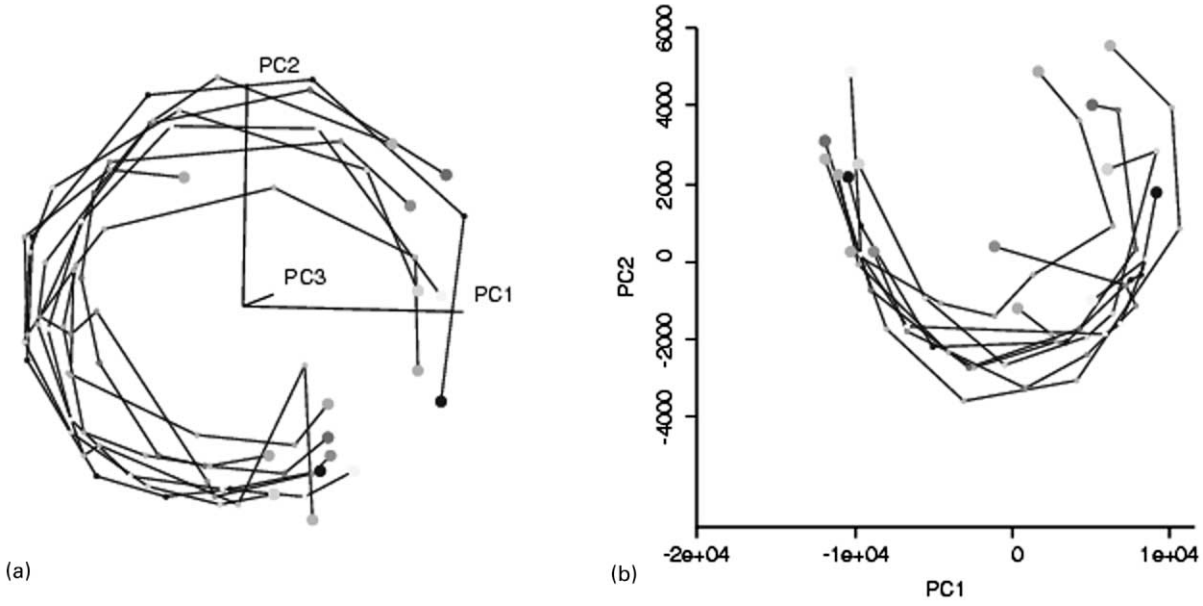

Fig. 10. Pose manifolds in PCA space. (a) Prototypes at $\theta = 90°$ and $\phi = [0, 10, \ldots, 90]$ projected onto first three principal components. (b) Prototypes at $\phi = 90°$ and $\theta = [60, 70, \ldots, 120]$ projected onto first two principal components.

distance. Images were always post-normalised by subtracting the mean intensity from each pixel and dividing by the intensity standard deviation.

7. Filtering for pose discrimination

Consider what sort of image filters would be appropriate for discriminating different poses. It is expected that different image features are important at different poses and that those features will be oriented differently. For example, the mouth may be important at frontal poses and the nose at profile poses. Therefore filters that highlight oriented features are appropriate. In this section, we investigate whether Gabor filters are useful for discriminating pose.

Gabor filters are oriented sinusoidal filters modulated by a Gaussian envelope. Examples of Gabor filters are shown in Fig. 7 for angles 0, 30, 60, 90, 120 and 150°. The real and imaginary parts are shown on the left and right, respectively. These filters have a natural application for pose estimation because pose estimation involves variations in orientation [7]. Freeman and Adelson have used separable oriented steerable filters for phase analysis, adaptive filtering, edge detection and shape from shading [6].

To see whether Gabor filters are useful for discriminating pose, let us evaluate the pose similarity ratio of Eq. (7) at a fixed pose but with varying Gabor filter orientation. The filter orientation γ is varied from 0 to 180° in 9° increments. The tilt angle is fixed at 90° (frontal view) and is not varied in the calculation of the ratio, i.e. $\delta\theta = 0°$. The yaw neighbourhood $\delta\phi$ is set to 30° and the size of the filters is 13×13. The result is a series of ratio values versus filter orientation. The process has been repeated at different fixed poses with yaw varying over the range [0°,90°] and tilt fixed at 90°. The results are shown in Fig. 8(a). Clearly the ratios vary smoothly with filter orientation and there are well-defined minima in the curves. The implication is that Gabor filters reveal oriented features in the facial images that are specifically appropriate for discrimination at a given pose.

In Fig. 8(b), the correlations between filter orientations and pose variations in tilt are presented. Yaw is fixed at 90° and the pose neighbourhood is $\delta\phi = 0°$, $\delta\theta = 10°$. Tilt is varied over 60–120°. Again, it is observed that the ratios vary smoothly with filter orientation and that the curves contain well-defined minima. We can conclude that features at a specific orientation are important for discriminating poses. This raises the question: does the best filter orientation vary with pose, and if so, how does the orientation vary across the pose sphere?

Let us now proceed to examine the best single orientation-selective Gabor filter for each pose by minimising the pose similarity ratio at each pose. To determine the best filter size, the average minimum ratio for a range of filter sizes is shown in Table 1. It can be noted that the ratio decreases monotonically with the filter size. Taking 13×13 as the filter size and using a neighbourhood of $\delta\phi = 20°$, $\delta\theta = 10°$, the optimal filter orientations and corresponding ratios are shown in Fig. 9.

Examining Fig. 9(a), it is clear that different orientations are optimal for different poses. Taking into consideration that the pose database itself contains spatial and pose misalignments, the variations in filter orientation with pose angle are reasonably gradual. There is also a fair degree of symmetry in the orientations about central yaw $\phi = 90°$.

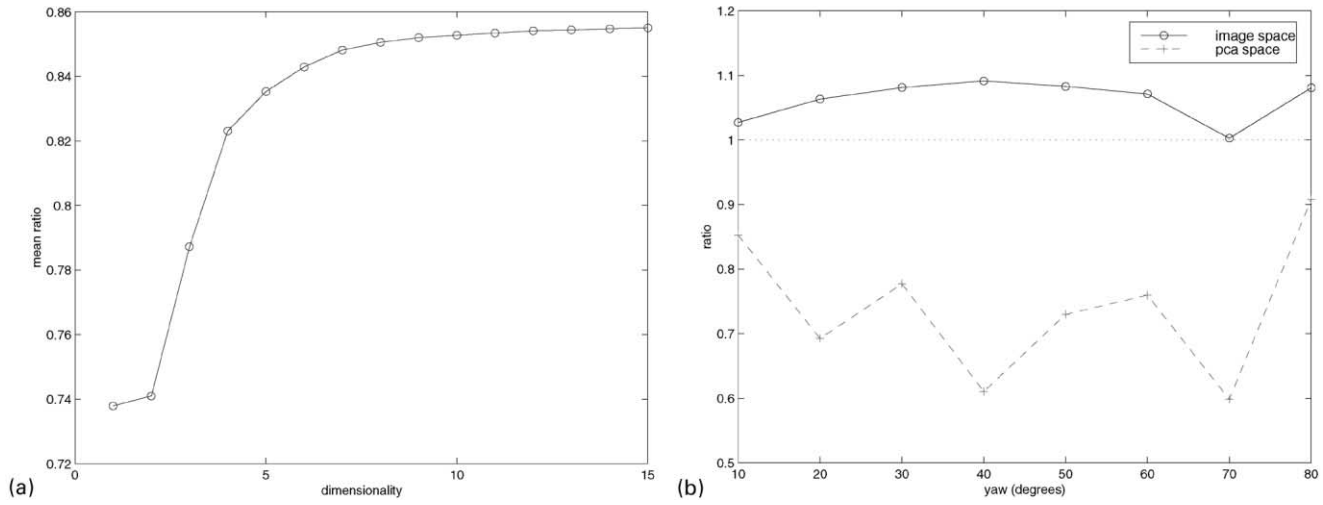

Fig. 11. Comparison of pose similarity ratios calculated in image space and PCA space for varying yaw angles. (a) Lowest ratio averaged over yaw versus number of PCA coefficients. (b) Ratios versus yaw angle for similarities calculated in image space and in PCA space using the first two coefficients.

In Fig. 9(b), the minimum ratios are represented as intensities with darker colours denoting lower (better) ratios. The pose angles containing a '×' have a ratio greater than 1. The results show that Gabor filters are able to discriminate faces from neighbouring poses except at some poses on the fringe of the pose sphere.

There are a few other points of interest from Fig. 9(b). The lowest ratios are at frontal yaw reinforcing the intuition that pose discrimination is easier at frontal views. The ratios when the subject is looking upwards are generally worse than when looking downwards. This could either indicate that the database acquisition system is less accurate at low tilts or it could be a natural phenomenon. The asymmetry in ratios about central yaw is due to misalignments and varying illumination conditions in the database.

To summarise, orientation-specific features are found in facial images at different poses and Gabor filters can be used to find these features. We now proceed to look at transformation for identity invariance.

8. Identity invariance through PCA

We have seen how orientation-selective filters can emphasise differences in pose, but can we also suppress identity? To obtain some invariance to identity, we investigate the use of PCA on the pose data. PCA is a linear transform based on Eigen Vectors, which are the orthogonal axes of maximum variance in the given set of data [1]. If the data lie in a linear sub-space of the original space, then a (usually

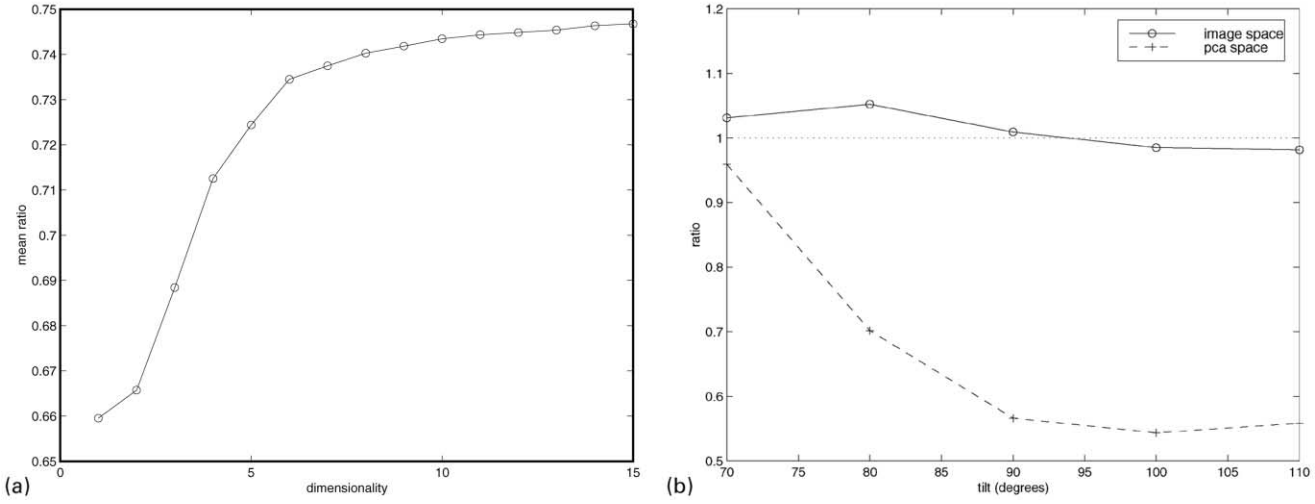

Fig. 12. Comparison of pose similarity ratios calculated in image space and PCA space for varying tilt angles. (a) Lowest ratio averaged over tilt versus number of PCA coefficients. (b) Ratios versus tilt angle for similarities calculated in image space and in PCA space using the first two coefficients.

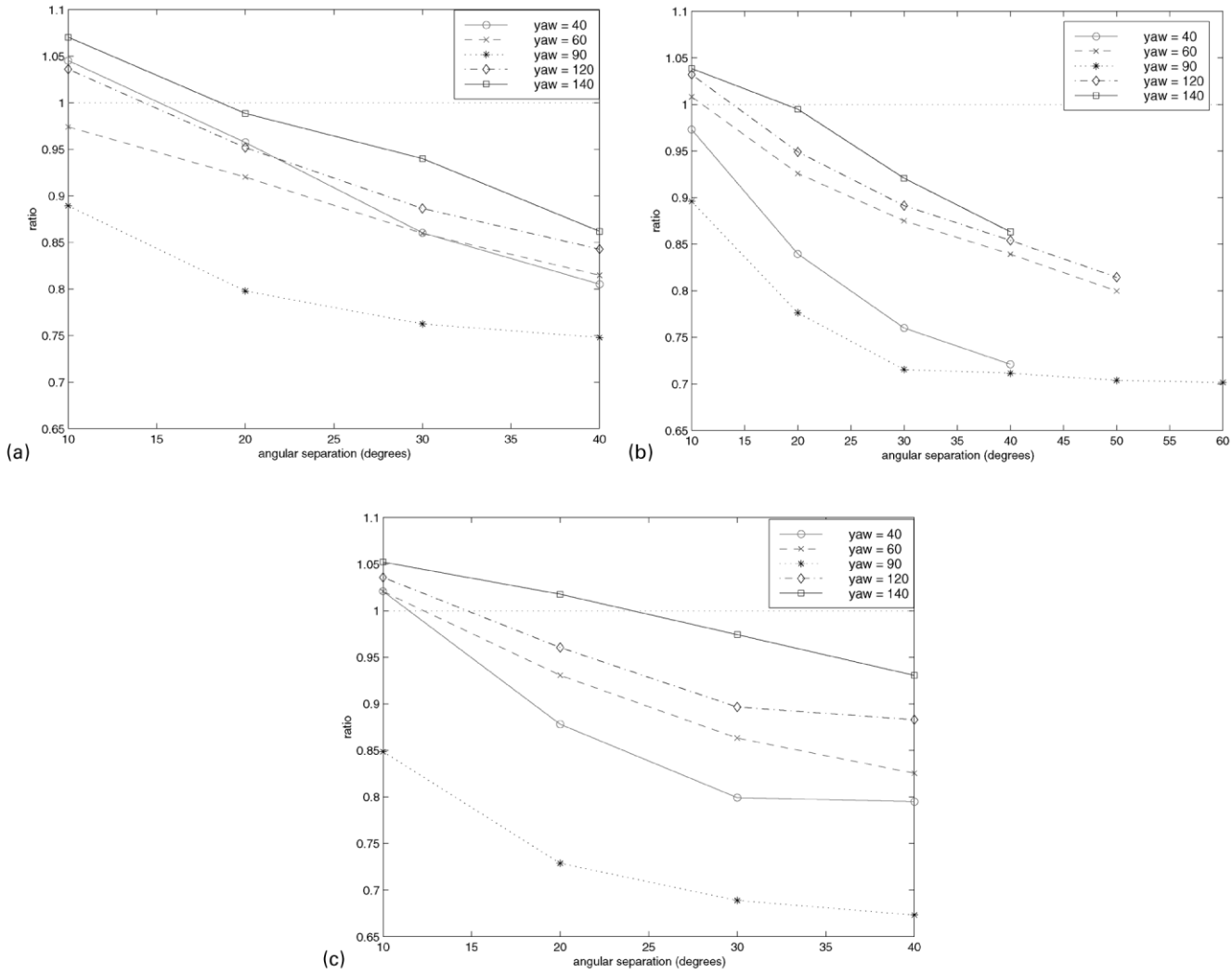

Fig. 13. Ratio versus different angular resolution across yaws at different tilts: (a) $\theta = 80°$; (b) $\theta = 90°$; (c) $\theta = 110°$.

significant) proportion of the Eigen Vectors will have only a small data variance in that direction. Therefore for the purposes of description of the data, only those Eigen Vectors with significant variance are required. Subsequently new data can be represented by its projection onto these significant Eigen Vectors resulting in a reduction in dimensionality. A previous investigation into pose distributions in PCA space found that continuous changes in yaw result in smooth manifolds in EigenSpace with identity collapsed [7]. Here we extend the study by calculating the pose similarity ratio based on similarities calculated in the PCA space.

To examine pose manifolds in PCA space, two PCA bases are calculated: one from images with tilt fixed at 90° and yaw varying from 0 to 90° and the other with yaw fixed at 90° and tilt varying from 60 to 120°. The range of poses for the investigation is restricted so that the PCA bases are based on a manageable range of intensity variations. In each case, prototypes from all eight subjects are used to construct a PCA basis and all images are blurred and normalised before use. Fig. 10 shows the prototypes of varying pose projected onto the first major principal components. Prototypes belonging to the same person are joined by a line in order of pose. In Fig. 10(a) for varying yaw, the curves form a horseshoe shape but the identities are clustered fairly tightly. The first two principal components account for 54% of the variance in the data. In Fig. 10(b) as tilt is varied, the same manifold shape is observed and the first two components account for 55% of the variance. The first two principal components largely describe changes in pose while the remaining components primarily encode changes in identity and facial expression. Therefore, projection onto the first two principal components provides a representation that is invariant to identity but sensitive to pose.

The PCA bases look appealing, but do they maintain sufficient discernibility between poses? To investigate, we calculate the pose similarity ratio *with distances calculated in PCA space*. The ratio is calculated at a range of poses covered by the PCA bases using a neighbourhood only in the axis of pose variation. For varying yaw, the neighbourhood is $\delta\phi = 10°$,

$\delta\theta = 0°$, and for varying tilt the neighbourhood is $\delta\phi = 0°$, $\delta\theta = 10°$. The average best ratio is plotted versus the number of principal components used where the average is over varying yaw in Fig. 11(a) and over varying tilt in Fig. 12(a). Comparing with the mean ratios for Gabor filters in image space shown in Table 1, the PCA-based ratios are much lower. Therefore PCA not only maintains good pose discrimination, it does so much more effectively than in the image space.

Using only the first two principal components to calculate similarities, the ratios are plotted for varying yaw and tilt in Figs. 11(b) and 12(b). On the same axes, the ratios are plotted for similarities measured in image space (no PCA, but blurred and normalised). Comparing the ratios with and without PCA, it is clear that PCA is a much more appropriate representation for the similarity calculations. Relating these results back to the Gabor filters, the non-PCA ratios plotted here are consistently higher than those obtained using Gabor filters, emphasising the need for Gabor filtering to exaggerate pose differences in image space.

In summary, PCA is an appropriate representation for pose similarity prototypes because it suppresses identity variations while maintaining sensitivity to pose. We have also seen that pose similarity ratios both in PCA space and after Gabor filtering in image space are better than those based on the original images. The fact that lower ratios are obtained with PCA than when using the Gabor filters does not necessarily mean the orientation-selective filters are no longer needed. Such a comparison is unfair because distance calculations are generally less robust in the high-dimensional image space due to the curse of dimensionality. Gabor filters are also expected to improve the smoothness of the PCA representation by reducing the sensitivity of the first two components to illumination changes.

9. Valid angular resolution

Logically there is a limit to the angular resolution with which poses can be discriminated using similarity-based methods. For example, at differences of 1° yaw, two 30×30 facial images would look so similar as to be indistinguishable. So the question arises: what is the minimum angular resolution at which pose differences can be discerned in the presence of varying identity and illumination? To find out, we modify the denominator of the pose similarity ratio. Eq. (9) becomes:

$$\bar{d}(\phi \pm \delta\phi, \theta \pm \delta\theta, f(\cdot)) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{y=\phi\pm\delta\phi}\sum_{t=\theta\pm\delta\theta} d(f(\mathbf{x}_{y,t}^{j}), (\mathbf{x}_{y,t}^{j}))}{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{y=\phi\pm\delta\phi}\sum_{t=\theta\pm\delta\theta} 1}$$

Here the ratio only involves neighbouring poses *at* $\phi \pm \delta\phi$ rather than all poses in the range of $\phi - \delta\phi, \ldots, \phi + \delta\phi$ and similarly for θ. This is akin to sampling the database at a lower angular resolution. Now we can plot the modified ratio versus angular resolution to find the minimum acceptable resolution.

At a range of yaws and two different tilts, the pose similarity ratio is calculated for the optimal filter (see Section 7) at varying yaw resolution $\delta\phi \in [10°, 60°]$ but with no tilt neighbourhood, $\delta\theta = 0$. The results are shown in Fig. 13 with the $r = 1$ threshold marked as a dotted line. As expected, the ratios monotonically decrease with angular separation because it is easier to discriminate larger changes in pose. For each tilt angle, 10° angular separation is not sufficient for some yaw angles because the ratio exceeds 1. At 20°, however, the angular separation is generally sufficient. The fact that the ratio is less than 1 at some yaws but greater than 1 at others implies that different angular resolution may be required at different poses. This requirement may arise because the problem is harder at these poses or because the noise in the acquisition system is higher at these poses.

Assuming that these results are indicative of the whole pose sphere, we can conclude that the greatest lower bound on discernible angular resolution is about 20°. Notwithstanding the minimum angular tolerance may be 10° or less at some poses.

10. Discussion and conclusions

We have presented an analysis of face similarity distributions under varying head pose for different types of image transformation with the aim of understanding pose in similarity space. The use of Gabor filters and PCA as transformations of prototype images to emphasise pose differences but suppress identity differences was examined. Orientation-selective Gabor filters were found to detect features for pose discrimination. Dimensionality reduction through PCA was found to provide invariance to identity while accurately describing pose changes. PCA also has the advantages of being understandable through visualisation and more computationally efficient since similarities are calculated in the low dimensional space. The lowest angular separation at which pose differences can be feasibly detected was also investigated. A greatest lower bound of approximately 20° was determined and the actual minimum resolution may be 10° or lower at some poses.

Overall, this work has shown that pose differences can be enhanced and identity similarities suppressed within a similarity-space framework using inexpensive algorithms. Such findings should facilitate the development of real-time pose estimation systems. Some remaining issues are: (i) The optimal filter orientation at each pose is not necessarily unique. Indeed, Gabor filters may not be the optimal filters for pose estimation. A more general approach could be taken by

adapting the filter orientation locally within the facial images [6]. (ii) The benefits of pose-selective filters and PCA can be combined. The main difficulty lies in creating PCA bases from images that have been filtered differently. (iii) PCA may not be the best linear projection for removal of identity information. For instance, linear discriminant analysis could be used to find the projection that maximises discrimination between faces at different poses. Such an approach has previously been adopted to achieve invariance to illumination conditions and facial expression [13].

References

[1] C. Bishop, Neural Networks for Pattern Recognition, Cambridge University Press, Cambridge, MA, 1995.

[2] S. Duvdevani-Bar, S. Edelman, A.J. Howell, H. Buxton, A similarity-based method for the generalisation of face recognition over pose and expression, IEEE International Conference on Face and Gesture Recognition, Nara, Japan, April 1998, pp. 118–123.

[3] S. Edelman, Representation is representation of similarities, Behavioral and Brain Sciences 21 (1998) 449–498.

[4] S. Edelman, Representation and Recognition in Vision, MIT Press, Cambridge, MA, 1999.

[5] K. Etemad, R. Chellappa, Discriminant analysis for recognition of human face images, Optical Society of America 14 (8) (1997) 1724–1733.

[6] W.T. Freeman, E.H. Adelson, The design and use of steerable filters, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (9) (1991) 891–906.

[7] S. Gong, S. McKenna, J. Collins, An investigation into face pose distributions, IEEE International Conference on Face and Gesture Recognition, Vermont, USA, October 1996, pp. 265–270.

[8] S. Gong, E. Ong, P. Loft, Appearance-based face recognition under large rotations in depth, Asian Conference on Computer Vision, Hong Kong, vol. 2, IEEE Press, New York, 1998 (pp. 679–686).

[9] S. Gong, E. Ong, S. McKenna, Learning to associate faces across views in vector space of similarities to prototypes, British Machine Vision Conference, Southampton, UK, vol. 1, 1998, pp. 54–64.

[10] N. Kruger, M. Potzsch, C. Von der Malsburg, Determination of face position and pose with a learned representation based on labelled graphs, Image and Vision Computing 15 (1997) 665–673.

[11] Y. Li, S. Gong, H. Liddell, Support vector regression and classification based multi-view face detection and recognition, Proceedings of the Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000, IEEE Computer Society, New York, 2000 (pp. 300–305).

[12] S. McKenna, S. Gong, Tracking faces, IEEE International Conference on Face and Gesture Recognition, Killington, Vermont, US, October 1996, pp. 271–277.

[13] B. Moghaddam, A. Pentland, Probabilistic visual learning for object representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 696–710.

[14] J. Ng, S. Gong, Performing multi-view face detection and pose estimation using a composite support vector machine across the view sphere, Proceedings of the IEEE International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time SystemsKerkyra, Greece, Kerkyra, Greece, IEEE Press, New York, 1999.

[15] E. Osuna, R. Freund, F. Girosi, Training support vector machines: an application to face detection, IEEE Conference on Computer Vision and Pattern Recognition, 1997, pp. 130–136.

[16] A. Pentland, B. Moghaddam, T. Starner, View-based and modular eigenspaces for face recognition, IEEE Conference on Computer Vision and Pattern Recognition, 1994, pp. 84–91.

[17] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C, 2nd ed., Cambridge University Press, Cambridge, MA, 1992.

[18] Y. Raja, S.J. McKenna, S. Gong, Tracking and segmenting people in varying lighting conditions using colour, Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 228–233.

[19] R.P.N. Rao, D.H. Ballard, Natural basis functions and topographic memory for face recognition, International Joint Conference on Artificial Intelligence, Montreal 1 (1995) 10–17.

[20] H. Rowley, S. Baluja, T. Kanade, Neural network-based face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1) (1998) 23–38.

[21] K. Sung, T. Poggio, Example-based learning for view-based human face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1) (1998) 39–51.

[22] R.Y. Tsai, A. versatile, camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses, IEEE Journal of Robotics and Automation RA-3 (4) (1987) 323–344.

[23] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.

[24] T. Vetter, T. Poggio, Linear object classes and image synthesis from a single example image, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 733–742.

[26] R.P. Würtz, Multilayer Dynamic Link Networks for Establishing Image Point Correspondences and Visual Object Recognition, Verlag Harri Deutsch, 1994.